

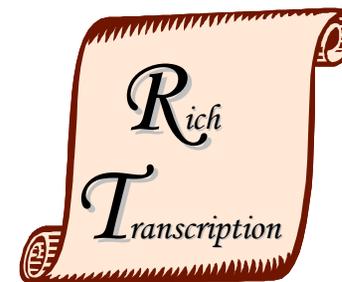
# RT-09 Speaker Diarization Results

2009 Rich Transcription Evaluation Conference

May 28-29, 2009

Melbourne, FL

Jérôme Ajot & Jonathan Fiscus



<http://itl.nist.gov/iad/mig/tests/rt/2009/>

# RT-09 Evaluation Participants

Site ID	Site Name	Evaluation Task			
		SPKR		STT	SASTT
		Audio	Audio/Video		
AMI	Augmented Multi-party Interaction: Univ. Sheffield, IDIAP, Univ. Edinburgh, Univ. of Technology Brno, Univ. Twente	X		X	X
I2R/NTU	Infocomm Research Site and Nanyang Technological University	X			
FIT	Florida Institute of Technology			X	
ICSI	International Computer Science Institute	X	X		
LIA/Eurecom	Laboratoire Informatique d'Avignon/ Ecole d'ingénieurs et centre de recherche en Systèmes de Communications	X			
SRI/ICSI	SRI International and International Computer Science Institute			X	X
UPM	Universidad Politécnica de Madrid	X			
UPC	Universitat Politècnica de Catalunya	X			

# Diarization “Who Spoke When” (SPKR)

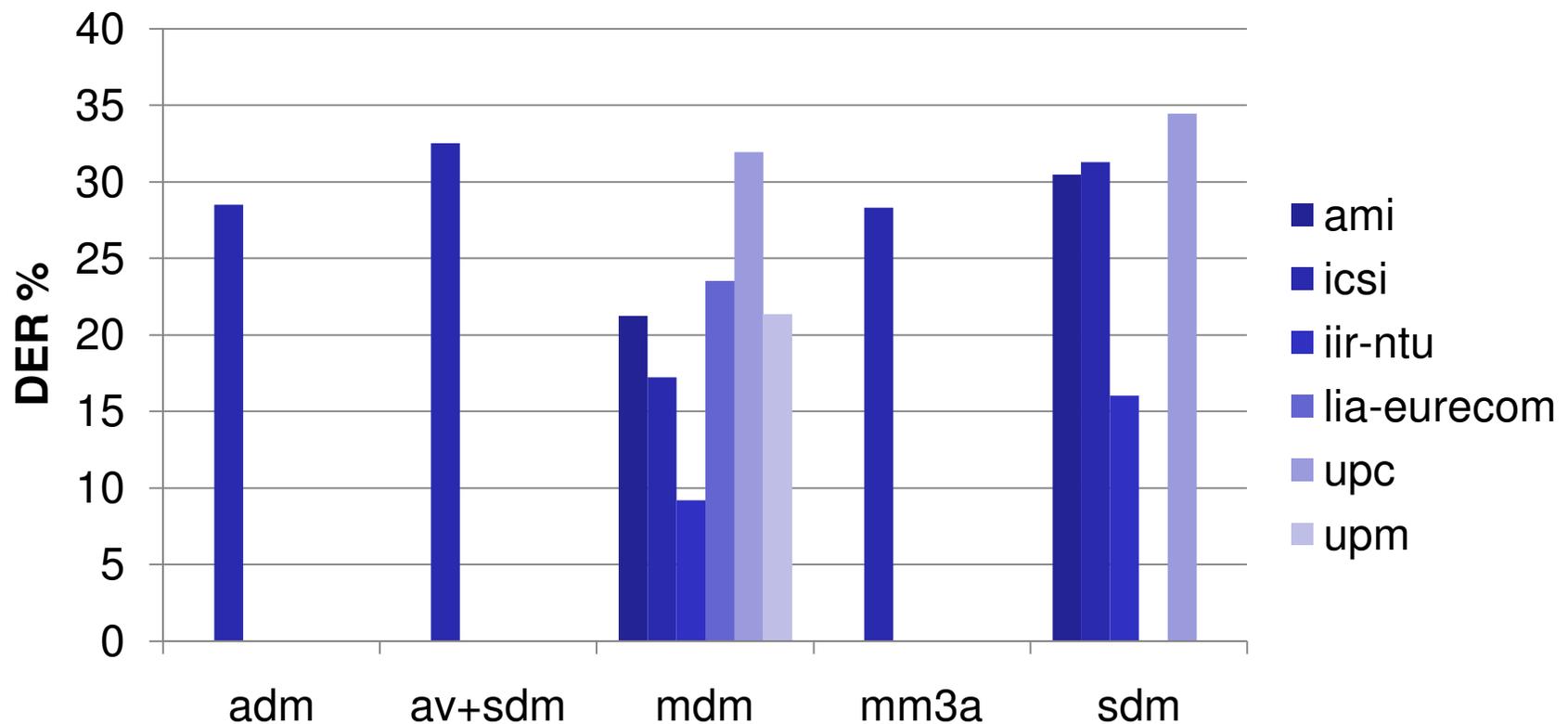
- Task:
  - Detect segments of speech and cluster them by speaker
- Primary input condition:
  - Multiple Distant Mics
- Participating sites:
  - AMI, IIR/NTU, ICSI, LIA/Eurecom, UPC, UPM
- Reference file construction: (not changed for RT-09)
  - Reference segment derived by:
    - force aligning the IHM audio to the reference transcripts using LIMSI tools
    - Segments built for each word were smoothed with a 0.3s window

# SPKR System Evaluation Method

- Step 1: Speaker alignment
  - A one-to-one mapping between reference speaker segment clusters and system determined speaker clusters
  - The mdeval tool was used with a +/- 250ms no-score collar around reference segment boundaries
- Step 2: Error metric computation
  - Diarization Error Rate (DER) – the ratio of incorrectly detected speaker time to total speaker time
  - Error Types:
    - Speaker assignment errors (i.e., detected speech but not assigned to the right speaker)
    - False alarms
    - Missed detections
  - Three scorings performed
    - All speech (Primary metric)
    - Non-overlapping speech (for backward compatibility)
    - Scoring as a Speech Activity Detection system

# RT-09 SPKR Results

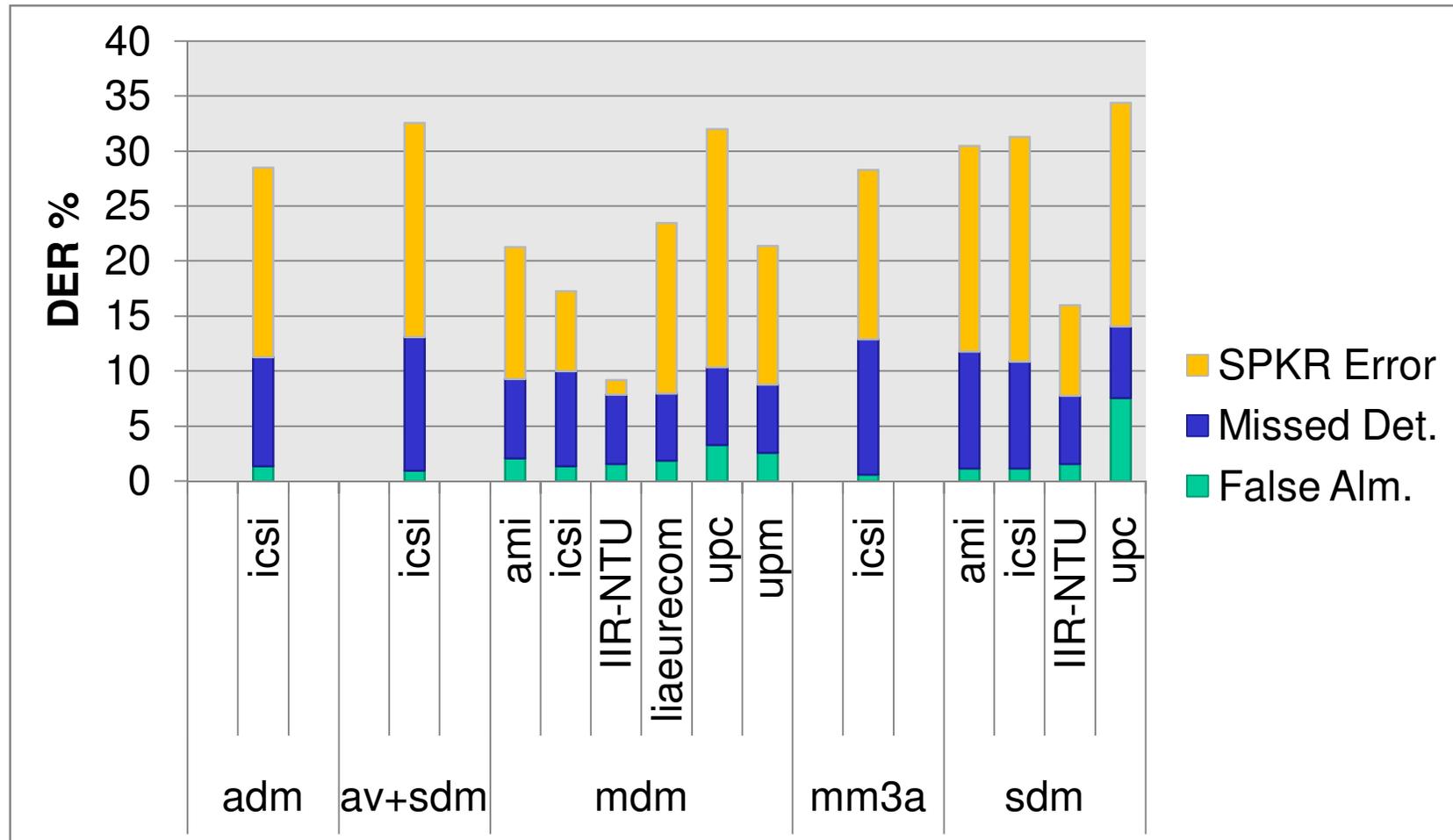
## Primary Systems, All Speech



- IIR-NTU has < 10%DER
  - But last test, it was ICSI
- Improvement with MDM < MM3A < SDM
- First use of video for diarization

# RT-09 SPKR Results

## Primary Systems, All Speech, Split by Error Type



- Speaker Error Dominates

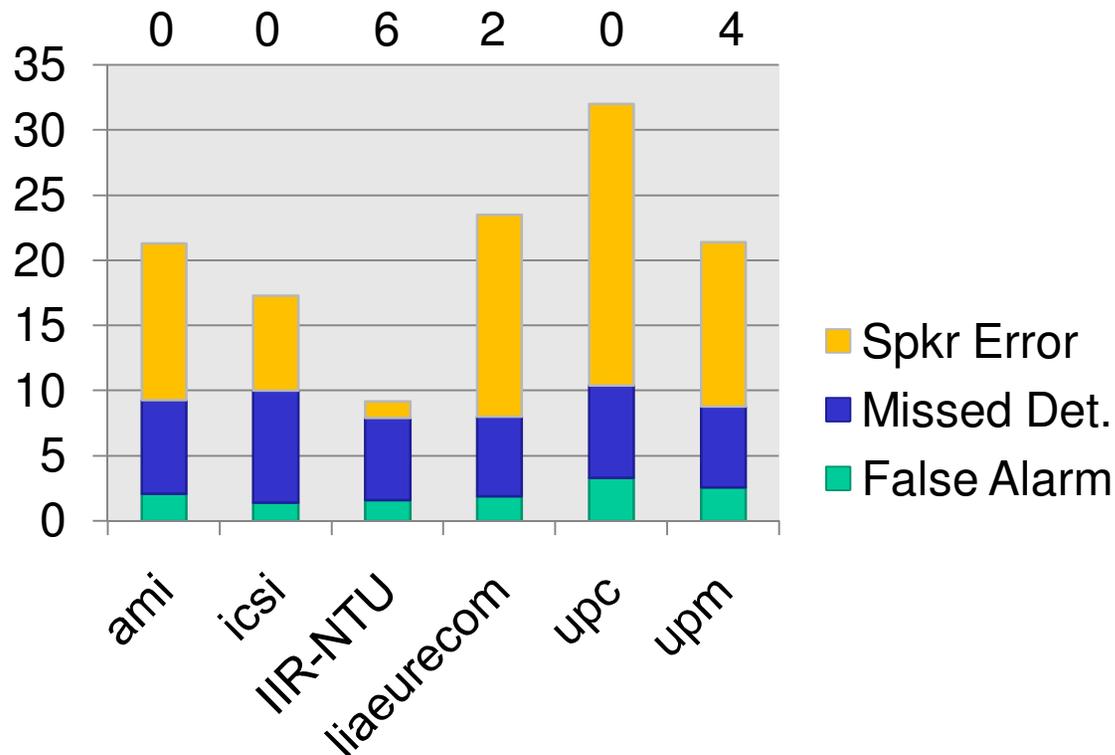
# MDM Detailed Analysis

- Focused analysis on MDM test condition
  - Correct detection of active speakers
  - All data vs. no overlapping speech vs. speech activity
  - DER variability by meetings
- Audio + Visual diarization
- Historical DERs

# RT-09 Primary SPKR MDM Systems

## DER Split by Error Type

Number of meetings with the correct # of speakers  
(out of 7)



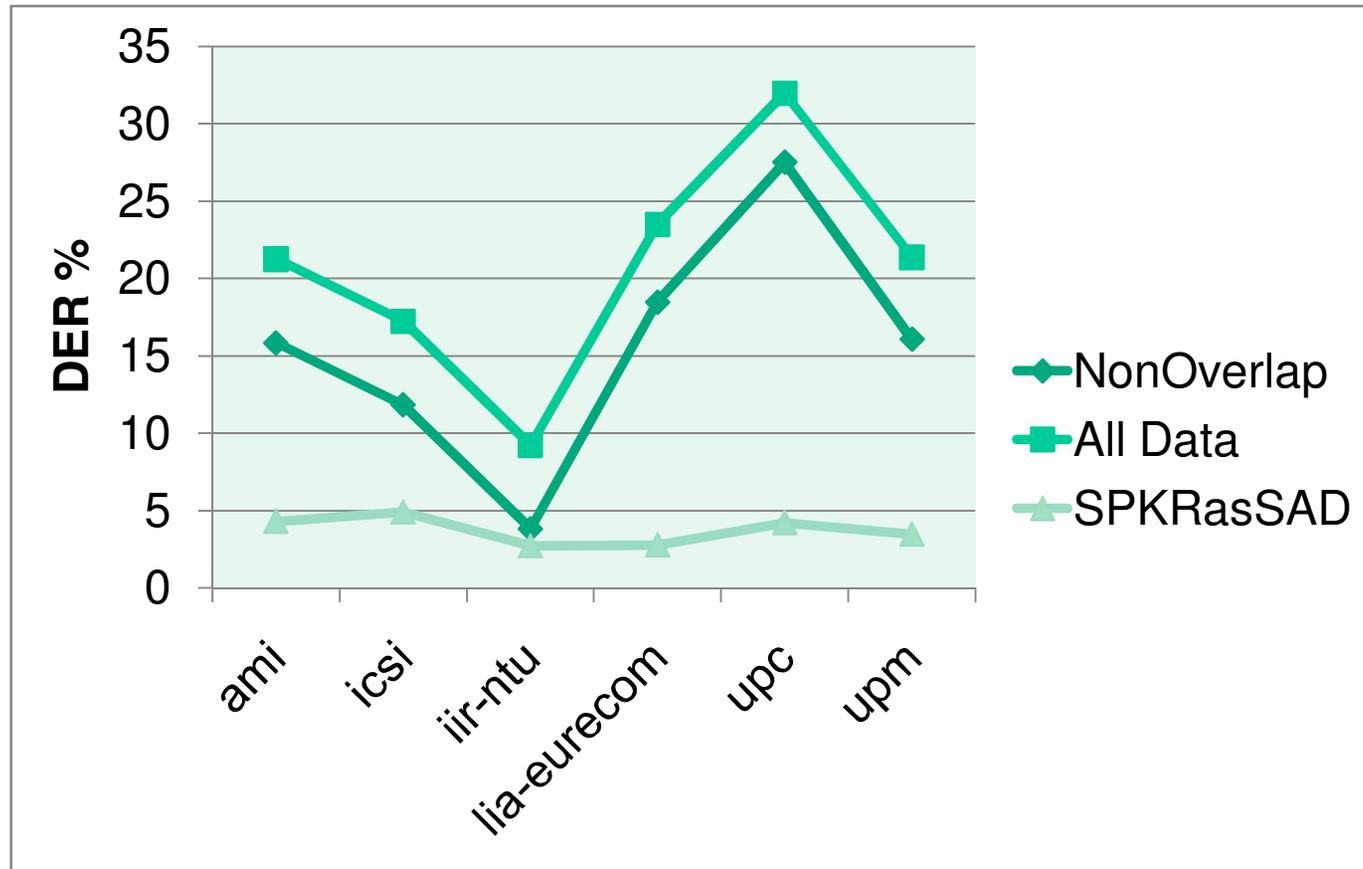
Questions:

Is there a meeting effect

- Speaker Errors dominate the scores, not for IIR-NTU
- False alarms and Missed Det. similar for all

# RT-09 SPKR Results

## Primary Systems, MDM Conference Data

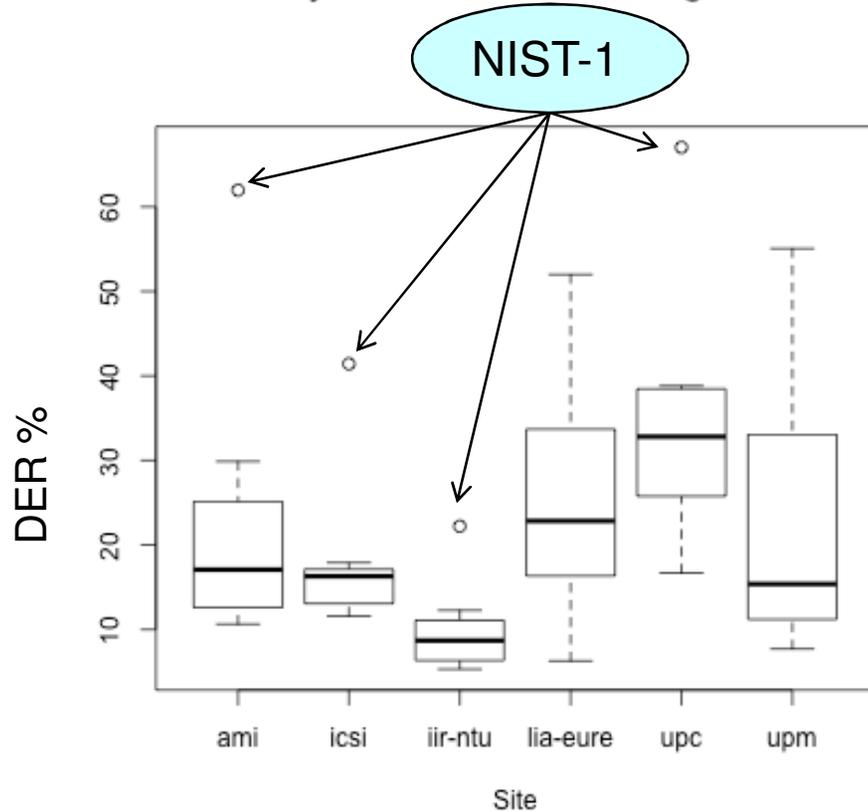


- High correlation between with/without overlap
- SAD scores are commensurate within domain

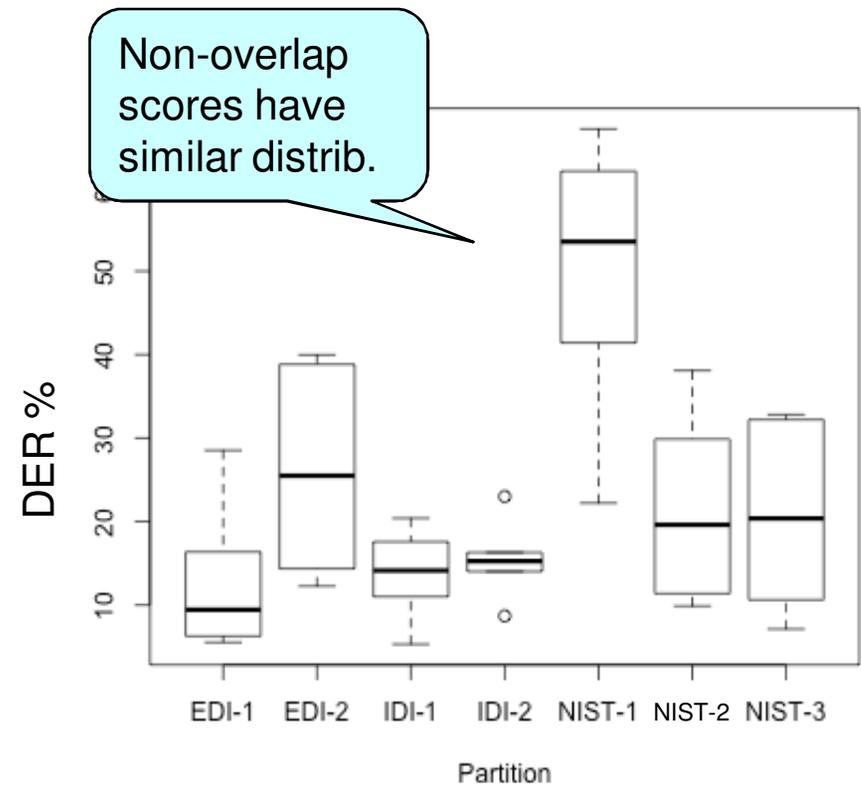
# RT-09 Primary SPKR MDM Systems

## Meeting DERs – within/across systems

DER by Site Across Meeting IDs



DER by Meeting IDs Across Sites

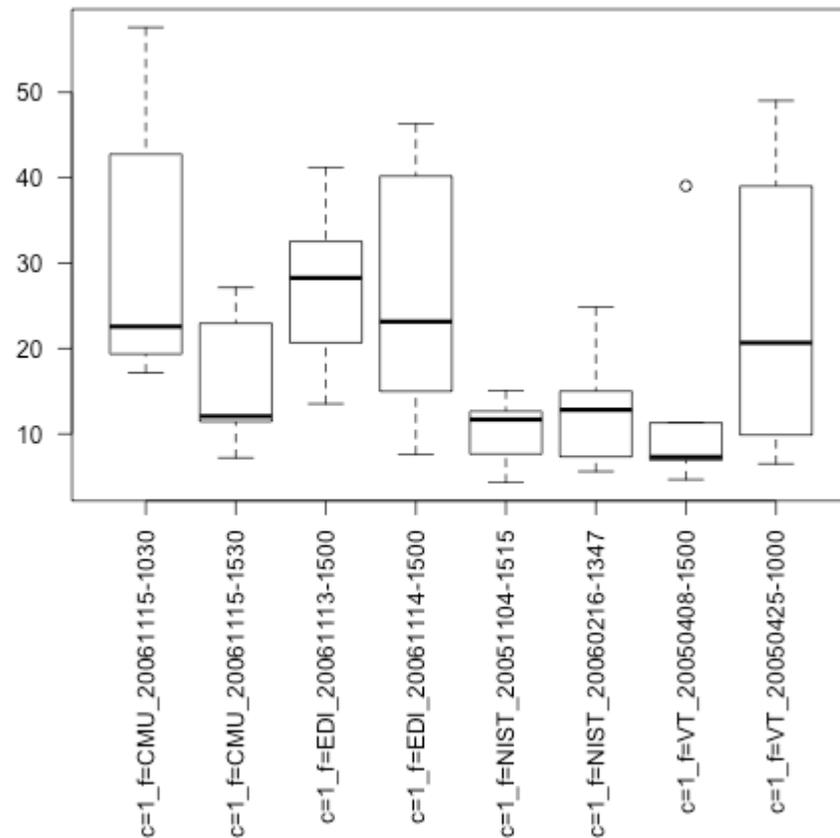


EDI_20071128-1000	EDI1
EDI_20071128-1500	EDI2
IDI_20090128-1600	IDI1
IDI_20090129-1000	IDI2
NIST_20080201-1405	NIST1
NIST_20080227-1501	NIST2
NIST_20080307-0955	NIST3

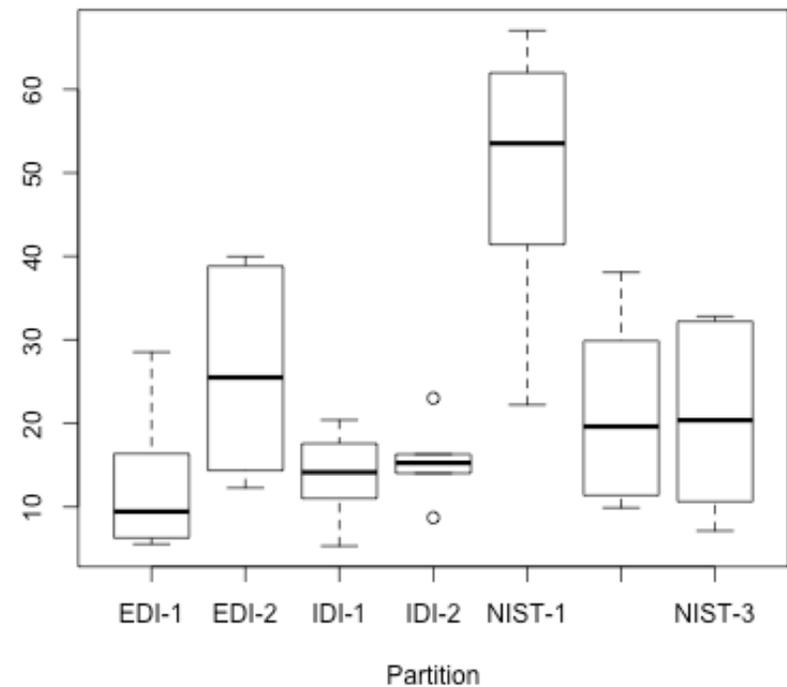
Small sample caveat: 6 systems / 7 meetings

# Meeting DERs: RT-07 vs RT-09

## RT-07

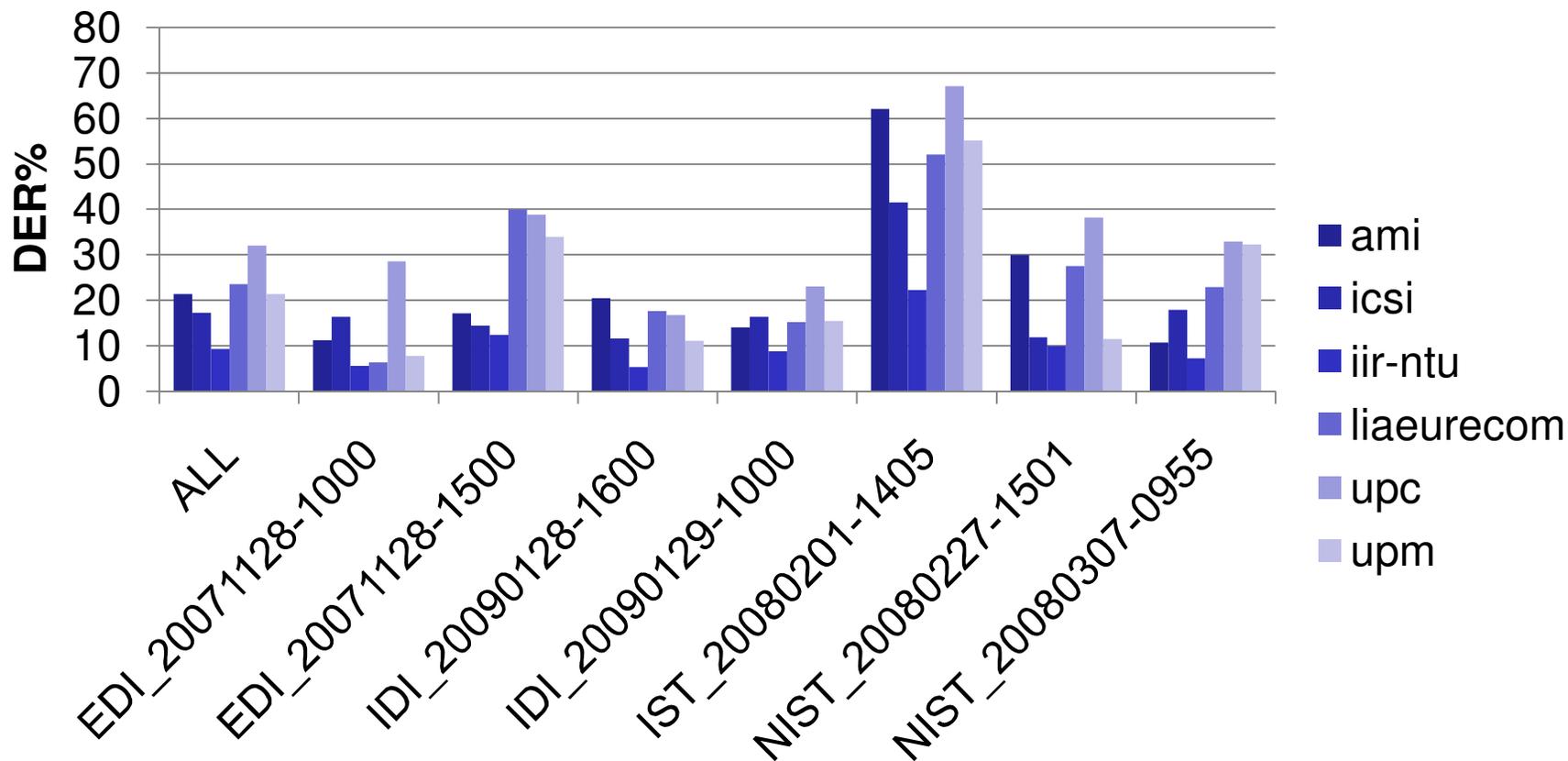


## RT-09

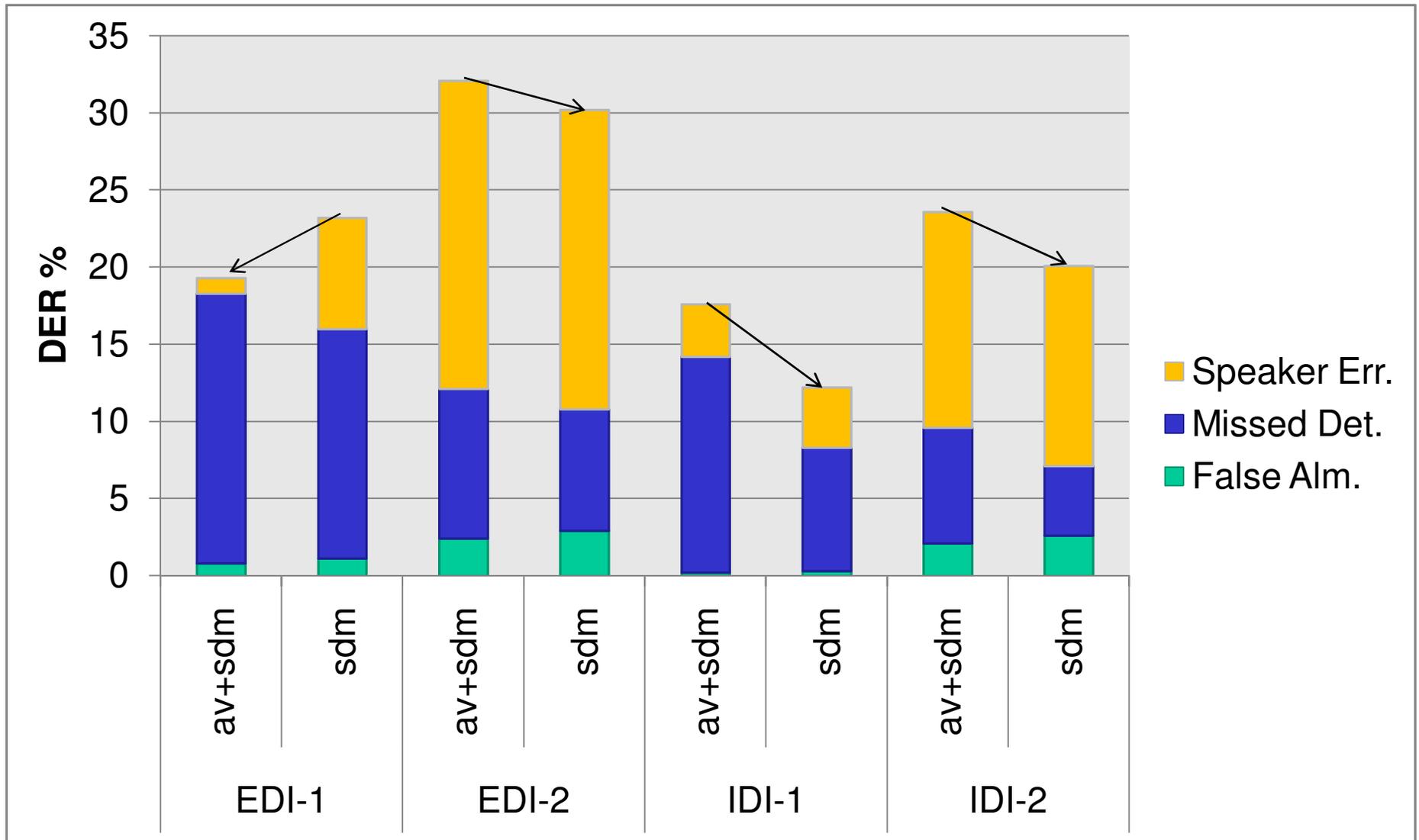


- Demonstrable meeting effect
- Large within meeting variation

# MDM Error Rates by Meeting

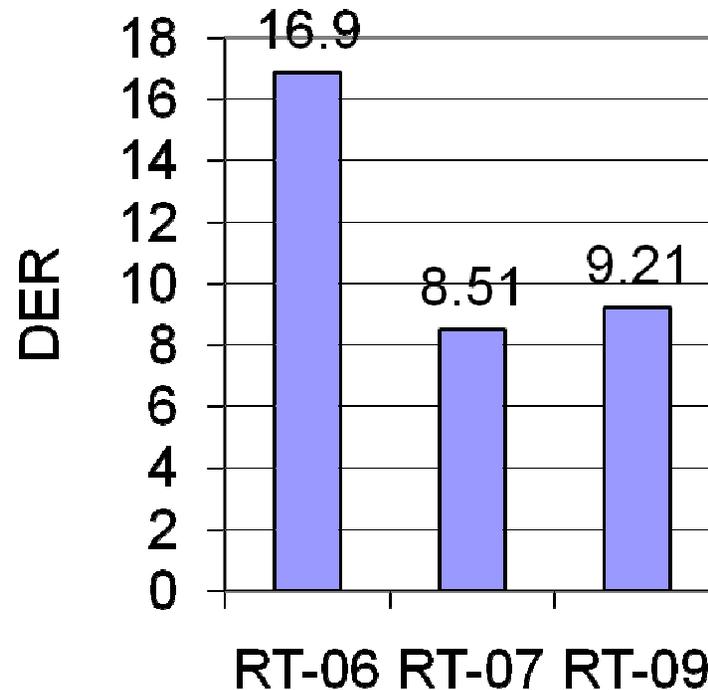


# ICSI SDM + Video Diarization



# Historical Best System MDM SPKR Performance

(Forced Alignment Mediated)



# Conclusions

- Bigger test sets are needed
  - The large variability in meeting error rates
- Like last year:
  - Lowest error rate system correctly detected the right number of speakers
- Has performance reach asymptote?
  - What the best performance you can get without solving overlap?